Australian Journal of Educational & Developmental Psychology Vol. 5, 2005, pp 91-106

Investigating Moderators of Test-retest Reliability in Screening Children's Cognitive Functioning[1]

Fiona H. Spencer
Queensland University of Technology

&

Laurel J. Bornholt
University of Sydney

ABSTRACT

This project examined some of the common constraints on reliable cognitive assessment for children over time. Repeated measures design (n = 133) was with younger (5 & 6 year old, n = 64) and older (7 & 8 year old, n = 69) children. Results showed that children's self-concepts moderated the test-retest reliability over extended intervals for younger and older children. Social self-categorizations moderated slightly the test-retest reliability for both age groups. For younger children only, personal self-categorization and self-reported talking about items moderated the reliability. In contrast, there were no moderators of test-retest reliability for shorter retest intervals. Other experiences, including recall, recognition and self-reported reflection on the content, did not moderate the test-retest reliability.

INTRODUCTION

From five- to seven-years of age are a productive time for children's cognitive development; it is traditionally an excellent time to analyse cognitive change (White, 1965), particularly because "children's thinking after age 7 seems to be different from their thinking before age 5" (in Sameroff & Haith, 1996, p. 3). It is well-known that children's cognitive functioning varies considerably with age to such an extent that younger and older children can seem quite distinct in terms of cognitive performance (White, 1996) and changes over time, particularly in cognition, are directly related to age (see Sameroff & Haith, 1996). In their classic work, Baltes, Reese and

[1] Contact
Dr Fiona H. Spencer (author for correspondence)
Lecturer
School of Learning and Professional Studies
Queensland University of Technology
Victoria Park Rd
Kelvin Grove Qld 4059 Australia
Email: f.spencer@qut.edu.au
Ph: 07 3864 3934
Fax: 07 3864 8265

Laurel J. Bornholt
Senior Lecturer
School of Development and Learning
University of Sydney
NSW 2006 Austsralia
Email: l.bornholt@usyd.edu.au
Ph: 02 9351 2618
Fax: 02 9351 2606

Nesselrode (1977) argued that age could be seen as an independent variable that affects a dependent variable, such as test performance. In a study of 6 to 12 year olds, van den Burg and Kingma (1999) found that age contributed to test performance in neuropsychological assessment. Consequently, it is assumed that younger children's cognitive assessments are less stable because they are developing quickly.

In investigations of children's rapidly changing cognitive ability (Sameroff & Haith, 1996), age should be a major consideration. This means that age may provide a reference point for the analysis of change over time and it may influence test-retest reliability. Yet the evidence is somewhat contradictory. For example, Schuerger and Witt (1989) examined five intelligence tests from 34 separate studies in relation to test-retest reliability, and found that age had a significant effect on reliability. However, Kelley-Gomez (1999) found in five- to eight-year olds that age did not alter test-retest reliability of a computerised version of a reasoning ability test. In addition, Dyche and Johnson (1991) studied the test-retest reliability of auditory-verbal attention over a four-week interval and found that the noted practice effect was not related to age. The most significant finding of Schuerger and Witt (1989) was an increase in reliability with age when testing intellectual functioning. This argument led us to the main research question – Are factors that moderate test-retest reliability of cognitive assessments also distinct by age groups?

Understanding the test-retest reliability of assessments is critical in the selection of tests used in clinical practice. This is particularly important when the assessment outcomes are used in monitoring growth and recovery of children's cognitive functioning. This applies in educational interventions that support children with learning difficulties, as well as clinical programs following accident or injury. In other words, we need to account for any extraneous aspects that may routinely alter the reliability of assessment materials from one occasion to another. These extraneous factors may alter the outcomes of assessment, making them not as reliable over time. We need to determine whether test scores have changed over time, and identify the possible sources of the change (see Hannan, 2005). It is an essential first step to examine features of the instrument before examining features of the child in the testing situation. Standard psychometric methods are used to first check that any variations in test scores are not due to low test-retest reliability.

To identify additional factors that may impact on the reliability, Schuerger and Witt (1989) examined five intelligence tests from 34 separate studies. They found that age had a significant effect on reliability. Other variables that may also alter the reliability of cognitive assessment is the focus of the current investigation.

The SYSTEMS School-Years Screening Test for the Evaluation of Mental Status was used because it provides reliable and valid indicators of children's general cognitive functioning (see Bornholt, Spencer, Fisher & Ouvrier, 2004; Ouvrier, Hendy, Bornholt & Black, 1999). The current study investigates the moderators of the reliability. Previous studies have established the test-retest reliability of the SYSTEMS screening test, showing strong test-retest reliability and high cognitive functioning stability over varying test-retest intervals (see Spencer, Bornholt & Ouvrier, 2003). We proposed that other aspects of children's behaviour in the test-retest interval may also moderate the test-retest reliability. This means that individual children may show lower or higher cognitive functioning test scores, for instance, where performance is moderated by self concepts or recall of the items. Spencer and Bornholt (2003) suggest the particular aspects of self-knowledge that may moderate test-retest reliability. These include self-concepts that integrate aspects of performance, talent, effort and task difficulty, as well as children's self-categorizations about cognitive activities (see Bornholt, 2005a; Bornholt & Ingram, 2001). Conventions of psychological assessment (see Anastasi & Urbina, 1997; Sattler, 2001) suggest other experiences that may alter the test-retest reliability. This could entail personal experiences such as children's memory of items including recall and recognition, and thinking about the items, as well as social experiences such as talking about items during the retesting intervals.

Background to the Project

We take a socio-ecological approach in understanding children's cognitive development (see Bornholt et al., 2004; Bjorklund, 1995; Bowes & Hayes, 1999; Morrison, Griffith & Frazier, 1996) to focus on personal and social aspects of children's experiences that moderate reliability of cognitive assessments. In principle, children's cognitive functioning is considered more broadly than actual performance. We contextualised our understanding of reliability as the responsiveness to experience of children's cognitive assessments, over time (see Bornholt, 2005b).

Personal and Social Experiences

Recall and recognition. Early research by Carmines and Zeller (1979) suggests that memory from initial interviews may influence responses at subsequent interviews. There are many similar situations where children are assessed on repeated occasions, in clinical and educational settings. Clearly, children may recall or recognize items from previous testing sessions. This could alter their scores, and support (or constrain) the correspondence between test scores over time, for some children and not others.

Memory can be tested in the form of recall or recognition of events or information. We considered recall to involve two processes of retrieval where a cue is first recognized without prompting and can then be recalled at a later time (see Newcombe, 1996). It is clear that recall may be cued (specific external cue given) or free (prompted only by a general request). Bjorklund (1995) suggested that some younger children are susceptible to suggestion in cued recall and may provide inaccurate information, so free recall is utilized in the current study. Recognition is conceptualized here as a single process, which does not require generation of a full response, only a decision concerning the accuracy of the response (Ellis & Hunt, 1989). Recognition occurs in tests when children can distinguish between old and new items (Bjorklund, 1995). Development of recognition with age is of particular interest. Gross and Hayne (1999) found that five to six year old children could recognize and describe their own drawings after a one-year delay. Although Bjorklund (1995) suggests that recognition ability appears to change little with age, work by Markham, Howie and Hlavacek (1999) suggests that younger children obtain lower scores of recognition than older children.

Learning through reflection. Actively thinking and talking about information within testing intervals may moderate test reliability over time. Bjorklund (1995) supported the notion that experience of a previous performance influences a second performance. Alternatively, Nelson (1996) argues that changes in memory reflect changes in cognition through the use of language as a cognitive tool. This means that actively thinking and talking about experiences may influence test scores over time and enable better performance.

The impact of time intervals. If children remember the items due to personal and social experiences, then they would perform better on the test when re-administered after a short interval. Anastasi (1988) suggests that four weeks is a satisfactory retest interval. She argues that if retest intervals are short, then more information would be recalled correctly. It is hypothesized that increases in test information through recall, recognition, talking, and thinking about the items would increase the test reliability. So, it is important to establish whether children's memory and experience in a short or long retest interval alters test reliability. In clinical testing, appropriate test-retest intervals would minimize score interference due to the memory of items. We therefore compared short retest intervals of two weeks with the four-week interval suggested by Anastasi, and a longer 12-week interval for the effects of memory and experience of items on test reliability.

Personal and social aspects of the self. Self awareness is an important consideration in the development of children's cognitive functioning (see Spencer & Bornholt, 2003). We examined effects of children's self-concepts of their cognitive functioning (see Bornholt, Ouvrier, Black, & Hendy, 1999) as a moderator of test-retest reliability over time. Substantial research in related

fields (e.g., Bornholt & Piccolo, in press; Brake & Bornholt, 2004; Eccles, Wigfield, Harold, & Blumfield, 1993; Bornholt & Ingram, 2001) considers self-concept about a wide range of activities such as reading, drawing, the body and movement. For this project, reliable indicators of self concepts are provided by the ASK-KIDS Inventory that create and are created by children's participation in activities (Bornholt, 1997, 2004; 2005ab; Marsh, Debus & Bornholt, 2005). The ASK-KIDS model of self-concepts integrates aspects of the person in terms of performance, natural talent, effort, and task difficulty in relation to particular activities such as reading, number and drawing (Bornholt, 2005c), that is readily extended to other activities (see Brake & Bornholt, 2004). This model was applied in the present project to indicate children's self-concepts about cognitive activities. It was expected that children's cognitive self-concepts are also moderators of test reliability.

Particular personal and social self-categorizations were also explored that may vary test scores from one assessment to another and alter test reliability. From self-categorization approach, self concepts need to be considered in context, whether personal or social contexts (see Bornholt & Ingram, 2001; Bornholt & Piccolo, in press). Personal self-categorizations are within the person such as variations and commonalities across time and content, given a sense of individuality. For instance, thinking about ones performance at reading compared to drawing is personal categorization by content, and thinking about pervious and current social skills is a personal categorization over time. In addition, social self-categorizations rely on comparisons within and among many salient social groups such as gender, age and ability groups, given a sense of belonging socially. Recent research uses correlational and experimental evidence to show the self-categorizations that support and constrain children's self-concepts and participation in cognitive, social and physical activities (Bornholt, 2000; Bornholt & Ingram, 2001; Bornholt & Piccolo, in press; Eccles et al., 1993). For cognitive activities, the research suggests a social basis to children's self concepts about number activities, a stronger personal than social basis for drawing, and for children with reading difficulties (Bornholt & Ingram, 2001; Coleman & Bornholt, 2003). We therefore explored the effects of both personal and social self-categorizations about general cognitive activities on test-retest reliability.

Summary

The main question is whether characteristics of the children and their experiences within the testing interval alter the test-retest reliability of children's cognitive assessment over time for younger and older children. We considered the impact of memory and self-reported item reflection, self-concepts, personal and social self-categorizations on test-retest reliability and hypothesize that these aspects will moderate test-retest reliability.

METHOD

Research Design

We used a repeated measures design (Time A, Time B, Time C) to test children's cognitive functioning, recall, recognition, self-reported refection, self-concept and self-categorization. Time A provided a baseline indicator. At Time B, children were randomly allocated to retest intervals of two (Time B2), four (Time B4) or twelve weeks (Time B12). Assessment at Time C (the third testing session) used a standard interval of four weeks from Time B.

Research Sample

The children (n = 133) were aged between 5 to 8 years with a mean age of 6.6 (SD = 1.1) from primary schools in Sydney, Australia. For analysis purposes, the children were grouped as younger (5 & 6 year old, = 64) or older (7 & 8 year old, n = 69). The current sample consisted of young children who took part in a larger project, conducted by the Children's Hospital at Westmead, with 5 to 12 year old children (N = 399). The larger sample was drawn from six

schools randomly chosen from within a 20km radius of the Hospital from a variety of socio-economic areas. The ABS (2001) Socio-Economic Indices for Areas – Index of Education and Occupation (SEIFA-IEO) was used to select schools from high, medium and low areas, which was representative of the general Sydney population. Permission was obtained from school principals and parents and the response rate of 49% was considered satisfactory. Children up to the age of 8 years were selected for the current study, as past research has shown a ceiling effect at age 10 (Ouvrier, Goldsmith, Ouvrier, & Williams, 1993). There were no sample manipulations of the larger sample in selecting the current study sample. Parents were notified in permission letters that their child might be tested on two subsequent occasions. There were 256 children aged 5 to 8 years in the larger sample and only 52% of these were included in the current sample of 133. These children were randomly selected from the 256 Hospital study children with an equal consideration of gender to enable even numbers of boys and girls, 65 were boys and 68 were girls. The sample included children from medium socio-economic areas (55%), some high areas (38%) and low areas (7%). Preliminary analysis using MANOVA did not show significant difference between the means of the cognitive functioning scores for children by areas, $F(4, 260)$, $= .03$, $p > .05$, or by gender $F(2, 262)$, $= 1.42$, $p > .05$. So the samples were combined in subsequent analyses.

Sampling constraints. The entire study involved each child being tested on three occasions. Overall, there were 399 individual testing sessions with the maximum time span from the first (Time A) to the last occasion (Time C) of 4 months. The minimum time span from Time A to a Time C was 6 weeks. The testing took two years due to staggered testing and a few initial recruitment difficulties. Although the test retest schedule in the design was quite complex, there were few occasions that varied the test interval by more than 1 or 2 days. General practical constraints determined the sample size and the time taken to complete testing, for example sampling worked around school holidays, children being away from school on testing days and one child included in the project moved away from the school during the time span required for the study.

Procedures

The project was approved by University, Hospital and State Government Education Department ethics committees. After sample selection and permission procedures, standard testing procedures were followed for interviewing each child on every occasion incorporated evenly paced items, without feedback to the child about correct or incorrect responses, children were not told that they would be retested on another occasion. Children were told that there would be participating in some activities and that there were no right or wrong answers, just what they thought. Table 1 shows when materials were administered, and sub-sample sizes.

Table 1: Summary of variables, materials, sub-sample size and time of assessment

Variables	Sub-Sample Size	Time of Assessment	Materials
Cognitive Functioning	133	A, B & C	SYSTEMS
Memory (Recall & Recognition)	133	B & C	Pre & Post questions
Experience (Thinking & Talking)	133	B & C	Pre & Post questions
Self-concept	133	A, B & C	ASK-KIDS
Self-categorization	84	C	Post questions

Materials

Assessment of cognitive functioning. The School-Years Screening Test for the Evaluation of Mental Status (SYSTEMS) (Ouvrier, Hendy, Bornholt & Black, 1999; 2000) is a 46-item screening test of cognitive functioning for children. It takes about 10 minutes to administer verbally in a one-to-one interview situation. The child uses a worksheet for three questions. The test was administered according to the SYSTEMS form and manual (Ouvrier et al., 2000). Each item of the test is scored dichotomously (1) correct response and (0) incorrect response, and scores are summed. The research reports that for Australian school children aged 4 to 12 years SYSTEMS is internally consistent (alpha = 0.92) for 4 to 11 year olds, test scores are unbiased by gender, $F(1, 1011), = 0.75, p > .05$, and socio-economic indicators $F(2, 1010), = 0.16, p > .05$ with strong inter-rater reliability ($r = .94$) (see Ouvrier et al., 1999). Ouvrier et al. (1999) and Russell, Bornholt and Ouvrier (2002) also show strong correlations between test scores and full general cognitive assessments ($r_s = .88$ and .75, respectively) using the Stanford Binet (Thorndike, Hagan & Sattler, 1986) and the Differential Ability Scales (Elliott, 1990). The screening test examines aspects of cognitive functioning, including orientation, attention, calculation, reading, comprehension, copying, writing and memory. An example of the test content is the immediate and delayed recall of the placement of an object.

Additional items. For the current study, six false questions (different for each session B and C) were incorporated into the test to determine false answers at Time B and Time C. These items were integrated into the test at various points. There were three false items before Question 1 and one each after Question 16, after Question 29 and after Question 37. The false questions were distinct enough to be different from any previous items, for example at Time B in the spelling section children were asked to spell 'Tamagotchi'. This is quite a difficult and unusual question that they might remember from the previous administration. According to Markham et al. (1999), it is appropriate to determine accurate recognition scores by subtract the number of false items each student purported to remember from the number of overall recognized items, this approach was taken in the current study.

Assessment of item memory and reflection. Memory questions were used to determine the number of recalled items and then recognized items. To test for any content recall, pre-test questions were administered at Time B and Time C prior to the administration of the cognitive functioning test. When testing recall, children were initially asked, "What questions were you asked last time?" If any items were recalled the child was classified as having recall of items.

To test for any content recognition, while the Time B and C tests were being administered, children were asked after each item, "Do you remember this question from last time, or not?" The response of yes or no was noted.

Assessment of personal and social reflections. At Time B and Time C, during the administration of the pre-test questions children were asked about their personal and social item reflections. The children were asked, on a scale from 1 to 5 (1 being not at all), how much they thought about and talked about the item/s in between the two testing intervals.

Children's self-concept about cognitive functioning. Cognitive self concepts were adapted from the ASK-KIDS Inventory (Bornholt, 1997; 2005a). The ASK-KIDS Inventory includes children's self evaluations about ten common activities for children (reading, number, drawing, movement, body, appearance, individuality, belonging, friends and communication). Responses use a five dot-point rating scale that is easy for young children to use. Research showed reliable self concepts based on conventional indicators of internal consistency (alpha coefficients > .70) and good fit of the model to children's responses (Adj *GFI* > .95; *ChiSq/df ratio* < 2.0; *RMSEA* < .05) using confirmatory factor analyses (see Bornholt, 1997, 2005a; Bornholt & Ingram, 2001; Marsh, Debus & Bornholt, 2005).

After administration of the cognitive screening test children were asked five self concept items (e.g., 'How good are you at these activities?', 'How difficult are these activities for you?'.

Responses were averaged (with reversed responses such as difficulty) to form self concepts from (1) low to (5) high.

Self categorizations about cognitive activities. Personal and social self-categorizations questions were asked of each child at Time C. Personal self-categorization addressed comparisons over time, 'How good were you at these activities this time as compared to last time?'. Social self-categorization addressed comparisons within age groups, 'How good were you at these activities compared to other kids your age?'. Responses using the five dot-point ratings were from (1) low to (5) high.

Self categorizations about cognitive activities. Personal and social self-categorizations questions were asked of each child at Time C. The personal self-categorization addressed comparisons over time, "How good were you at these activities this time as compared to last time?". Social self-categorization addressed comparisons within age groups, "How good were you at these activities compared to other kids your age?" Responses using the five dot-point ratings were from (1) low to (5) high.

Analyses

The cognitive functioning test-retest correlations were computed using Pearson Correlation Coefficients for each group. Group comparisons included: recall of items (no and yes); recognition of items (low and high median split); thinking about items (no and yes); talking about items (no and yes); self-concepts (low and high by median split) and self-categorizations (low and high median split). Using a Fisher r-to-z transformation (Freed, Ryan & Hess, 1991, Vassar Stats web site http://faculty.vassar.edu/lowry/rdiff.html) the two person correlations were compared for each area of interest. A probability level of $p<.05$ was considered as significant for this study. A non-directional two-tailed test of significance was used in all analyses where H_o does not equal H_1.

The main question was whether personal and social factors alter the test-retest reliability. Moderator variables control the statistical properties of the test is determined using the pearson correlation coefficients (reliability results) and the Fisher r-to-z transformation, which can provide a comparison that can determine statistically significant differences between two reliability results.

RESULTS

Recall and Recognition

With regard to the number of items each child recognized, the six false items were incorporated within the test to indicate false answers and to calculate an appropriate recognition score. Table 2 reports the frequency of false answers for all children. A t-test showed a significant difference between the means for the number of false answers at Time B and Time C $t(131) = 1.92$, $p > .05$ (two tailed), $d = 0.16$. According to Cohen (1988) this shows a small difference between the two sample means being compared (Huck, 2004). The proportion of children giving false answers increased from 43% at Time B to 59% at Time C. Chi square results showed no significant difference between the model (50% of children giving no false answers, 30% giving one or more and 20% giving 2 or more false answers) and the sample (Time B χ^2 (2, N = 133) =3.34, $p = >.05$ and Time C χ^2 (2, N = 132) = 5.29, $p > .05$).

In determining accurate recognition scores, for each child the number of false answers was subtracted from the number of recognized items reported at each testing occasion (Markham et al., 1999). The results of Fisher r-to-z transformation in Table 3 and 4 showed that recognition and recall did not affect the reliability of cognitive functioning over various time intervals.

Table 2: Frequency of false answers

Number of False Answers	Frequency at Time B	Frequency at Time C
0	76	54
1 only	29	46
2 or more	28	32
Missing	0	1
Total	133	132

Table 3: Fisher r-to-z transformations of recall and recognition groups over varying intervals for younger children

Variables	Groups	Time								
		A to B2			A to B4			A to B12		
		r	z	p	r	z	p	r	z	p
Recall B	Nil	.93 (7)	1.37	ns	.85 (14)	-.24	ns	na		
	Yes some	.82 (14)			.88 (9)					
Recognition B	Low(Mdn <25)	.95 (7)	1.10	ns	.87 (12)	0.00	ns	.88 (13)	.77	ns
	High(Mdn ≥25)	.83 (14)			.87 (11)			.95 (7)		
		B to C			A to C					
		r	z	p	r	z	p			
Recall C	Nil	.88 (38)	-.79	ns	.88 (38)	.31	ns			
	Yes some	.92 (26)			.86 (26)					
Recognition C	Low(Mdn <34)	.86 (32)	1.13	ns	.82 (32)	1.20	ns			
	High(Mdn ≥34)	.92 (32)			.90 (32)					

sample number in brackets, na = not available due to small numbers

Table 4: Fisher r-to-z transformations of recall and recognition groups over varying intervals for older children

Variables	Groups	Time								
		A to B2			A to B4			A to B12		
		r	z	p	r	z	p	r	z	p
Recall B	Nil	.93 (8)	.13	ns	.91 (11)	.49	ns	.83 (14)	-.81	ns
	Yes some	.92 (16)			.86 (13)			.93 (7)		
Recognition B	Low(Mdn<37)	.97 (7)	1.34	ns	.82 (11)	.78	ns	.89 (14)	.73	ns
	High(Mdn ≥37)	.87 (17)			.91 (13)			.76 (7)		
		B to C			A to C					
		r	z	p	r	z	p			
Recall C	Nil	.85 (22)	-.28	ns	.85 (22)	-.14	ns			
	Yes some	.87 (47)			.86 (47)					
Recognition C	Low(Mdn <43)	.88 (31)	.33	ns	.85 (31)	0.00	ns			
	High(Mdn ≥43)	.86 (38)			.85 (38)					

sample number in brackets

Self-Reported Thinking or Talking About Items

Many children self-reported thinking about the items between Time A and Time B, (45%) and between Time B and Time C (58%). Also, 22% and 35% of children reported talking about the items between Time A and B, and Time B and C respectively. Individual variations in the extent to which children self-reported to think or talk to other people about the items within the testing intervals did not impact on the test-retest reliability of the screening test, except for younger children talking about the items, see Figure 5.

Self-concept and Self-categorization

Moderation of the test-retest reliability by children's self-concepts is shown in Tables 7 and 8. The results show three significant differences in test-retest reliability between Time A and Time B2 (two week delay), between Time B to Time C and between Time A to C. Although the test-retest reliability remains strong for each of the sub-groups investigated, it appears that test-retest reliability is reduced where children compare their cognitive assessments to how good they are when comparing two time testing sessions. Those with high self concepts overall tended to obtain lower reliability results than those children with lower self-concepts.

Table 9 and 10 reports the Fisher r –to-z transformations of personal and social self-categorizations. The results show a significant difference in test-reliability results for social comparisons made between Time B and Time C for younger and older children. Specifically, those children with low social self-categorizations tended to have lower test-reliability results than those with high levels of self-categorizations. The results also show a significant difference in test-reliability results for personal comparisons for younger children.

Table 5: Fisher r-to-z transformations of self-reported thinking and talking groups over intervals, for younger children

Variables	Groups	Time								
		A to B2			A to B4			A to B12		
		r	z	p	r	z	p	r	z	p
Think B	Nil	.82 (10)	-1.52	ns	.77 (8)	-.75	ns	.88 (12)	-1.28	ns
	Yes some	.96 (11)			.89 (15)			.97 (8)		
Talk B	Nil	.90 (17)	.05	ns	.86 (10)	-.13	ns	na		
	Yes some	.89 (4)			.88 (7)					
		B to C			A to C					
		r	z	p	r	z	p			
Think C	Nil	.91 (21)	.19	ns	.85 (21)	-.58	ns			
	Yes some	.90 (43)			.89 (43)					
Talk C	Nil	**.87 (44)**	**2.63**	**.00**	.87 (44)	-.67	ns			
	Yes some	**.97 (20)**			.91 (20)					

sample number in brackets, na = not available due to small numbers, bold = significant differences found

Table 6: Fisher r-to-z transformations of self-reported thinking and talking groups over intervals, for older children

Variables	Groups	Time								
		A to B2			A to B4			A to B12		
		r	z	p	r	z	p	r	Z	p
Think B	Nil	.94 (10)	0.00	ns	.82 (12)	-.79	ns	.85 (15)	-.26	ns
	Yes some	.94 (8)			.91 (12)			.89 (6)		
Talk B	Nil	.93 (20)	-.08	ns	.83 (17)	-.83	ns	.75 (16)	-.59	ns
	Yes some	.94 (4)			.93 (7)			..89 (5)		
		B to C			A to C					
		r	z	p	r	z	p			
Think C	Nil	.86 (35)	-.33	ns	.86 (35)	.15	ns			
	Yes some	.88 (34)			.85 (34)					
Talk C	Nil	.90 (42)	1.44	ns	.88 (42)	1.17	ns			
	Yes some	.80 (27)			.79 (27)					

sample number in brackets

Table 7: Fisher r-to-z transformations of self-concept groups over intervals, for younger children

Variables	Groups	Time								
		A to B2			A to B4			A to B12		
		r	z	P	r	z	p	r	z	p
Self-Concept B	Low(Mdn <4.6)	.94 (7)	1.0	Ns	.92 (11)	1.22	ns	.91 (11)	0.00	ns
	High(Mdn ≥4.6)	.82 (14)			.76 (12)			.91 (9)		
		B to C			A to C					
		r	z	P	r	z	p			
Self-Concept B	Low(Mdn <4.6)	**.94 (29)**	**1.96**	**.02**	.85 (29)	-.45	ns			
	High(Mdn ≥4.6)	**.84 (35)**			.88 (35)					
Self-Concept C	Low(Mdn <4.8)	.90 (31)	.68	Ns	**.96 (31)**	**2.48**	**.00**			
	High(Mdn ≥4.8)	.86 (33)			**.86 (33)**					

sample number in brackets, bold = significant differences found

Table 8: Fisher r-to-z transformations of self-concept groups over intervals, for older children

Variables	Groups	Time								
		A to B2			A to B4			A to B12		
		r	z	P	r	z	p	r	z	p
Self-Concept B	Low(Mdn <4.2)	.95 (9)	.61	ns	.82 (12)	.67	ns	.75 (8)	1.01	ns
	High(Mdn ≥4.2)	.91 (15)			.90 (12)			.91 (13)		
		B to C			A to C					
		r	z	P	r	z	p			
Self-Concept B	Low(Mdn <4.2)	.90 (29)	.84	ns	.85 (29)	.15	ns			
	High(Mdn ≥4.2)	.85 (40)			.86 (40)					
Self-Concept C	Low(Mdn <4.6)	**.91 (34)**	**1.70**	**.04**	.90 (34)	1.48	ns			
	High(Mdn ≥4.6)	**.80 (35)**			.80 (35)					

sample number in brackets, bold = significant differences found

Table 9: Fisher r-to-z transformations of personal and social self-categorization groups over intervals, for younger children

Variables	Groups	Time					
		B to C			A to C		
		r	z	P	r	z	p
Personal Self Categ.	Low(Mdn <5.0)	**.81 (10)**	**1.67**	**.04**	.88 (10)	.11	ns
	High(Mdn≥5.0)	**.95 (32)**			.89 (32)		
Social Self Categ.	Low(Mdn <5.0)	**.79 (10)**	**2.98**	**.02**	.85 (10)	.64	ns
	High(Mdn≥5.0)	**.96 (32)**			.91 (32)		

sample number in brackets, bold = significant differences found

Table 10: Fisher r-to-z transformations of personal and social self-categorization groups over intervals, for older children

Variables	Groups	Time					
		B to C			A to C		
		r	z	P	r	z	p
Personal Self Categ.	Low(Mdn <4.0)	.94 (61)	0.00	ns	.95 (61)	1.17	ns
	High(Mdn≥4.0)	.94 (23)			.91 (23)		
Social Self Categ.	Low(Mdn <4.0)	**.89 (26)**	**165**	**.04**	.92 (26)	0.00	ns
	High(Mdn≥4.0)	**.95 (58)**			.92 (58)		

sample number in brackets, bold = significant differences found

DISCUSSION

This research project examined the moderators of test reliability for the SYSTEMS cognitive screening test. The overall aim was to examine factors that are proposed to moderate the test-retest reliability of cognitive assessment for children. We included recall, recognition, self-reported thinking and talking about the items, self-concept, personal self-categorization and social self-categorization as possible moderator variables. In summary, there were two main moderators of test-retest reliability for assessment using SYSTEMS cognitive screening - children's self-concepts and social self-categorizations.

The main finding was that for seven- and eight-year old children cognitive functioning test-retest reliability was not moderated by many of the factors suggested in the literature. These included how much children remembered from previous testing and whether they thought or talked about the items. The key factors were self-concepts and social self-categorizations. This was similar for five- and six-year old children, for whom self-reported talking about the items and personal self-categorization also moderate the test-retest reliability.

Self concepts. How children evaluate themselves as cognitive self concepts moderates test-retest reliability over extended intervals. Younger and older children with reported low self-concepts had more stable test scores than children with high self-concepts whose test scores were more variable. It appears that the younger and older children with low-self concepts may have been under-estimating how good they were at the cognitive activities. Their scores increased over time, with even more of a difference between scores than children with high self-concepts. Perhaps not being told why they were being re-tested had an impact on their self-concepts. Children may have thought that they did not perform well the first time and needed to be given the test again. This suggests that we need to communicate the purposes of testing, and particularly re-testing. The older children's experiences suggest better understanding initially. Yet for the older children, this difference was also evident in the third testing session.

Self-categorizations. It seems that children's experiences of cognitive assessments are not necessarily congruent with test scores. As found in many other contexts (see Bornholt, 2005a; Brake & Bornholt, 2004), children may under- or over-estimate how good they are at cognitive assessments compared to other children their age. In principle, how children categorise themselves personally and socially underpins children's attitudes, intentions and behaviours (see Bornholt, 2005a). Recent research in educational and clinical settings shows individual variations and particular trends in such processes of self-categorization and justification (Bornholt, 2005b; Bornholt & Ingram, 2001; Bornholt & Piccolo, in press; Coleman & Bornholt, 2003). In this context, age was relevant to how children think about cognitive assessments. It is useful to consider these self-categorization processes as forms of self-stereotyping that are vital to what motives children to participate in activities (Bornholt, 2005c). In this context, retest reliability was higher and more stable over time for younger and older children with high social categorizations, than children with low social categorizations. In addition, personal self categorizations over time were also important in test-retest reliability for younger children. Constant attention to children's self categorizations highlight the key social-cognitive processes that are vital in each situation. We need to consider the role of content comparisons that children make, their ideas of traits that are relatively stable to varying over time, in addition to children's social categorizations in terms of ability, age and gender groups.

Conclusions

The results lead us to conclude that the SYSTEMS cognitive screening test is generally stable over time, and that we do not need to be unduly concerned about minor variations when using the screening test. The test can be applied particularly to situations that require close monitoring of children's cognitive functioning. It is important that children understand our reasons for testing and particularly for re-testing.

In the main, the results confirmed that SYSTEMS is a sound indicator of children's cognitive functioning, and test-retest reliability was not moderated by children's experiences in terms of memory (recall and recognition) and self-reported reflection (talking and thinking about items, except for younger children talking about the items). Further research is warranted to test whether self-reported reflection underestimate the amount children spoke about the items that may reflect a desire to do well or not appear to be cheating. The findings show that in general test-retest reliability was evident across people and situations. In contrast to early suggestions by Carmines and Zeller (1979) and Bjorklund (1995), it appears that with brief indicators of cognitive functioning, that are psychometrically sound such as the SYSTEMS screening, the role of children's experience of previous performance is not of major importance. This has significance for clinicians using cognitive tests over time to investigate changes in children's functioning.

Further research is therefore warranted with other cognitive tests to confirm the findings with SYSTEMS screening test. The project covers many possible factors that may have altered retest reliability. For now, the findings are specific to situations using SYSTEMS cognitive

screening. We can speculate that the findings would also apply to similar assessments with appropriate full cognitive and other assessments, although test length would be a confounding factor with younger children and in clinical settings. The findings make a valuable contribution to understanding the interesting variations and range of factors that may alter test-retest reliability for such brief cognitive assessments, and broaden the scope of applications in terms of individual qualities and situations that assessors routinely encounter in repeated cognitive assessments with children.

REFERENCES

Anastasi, A. (1988). *Psychological testing*. New York, USA: Macmillan.

Anastasi, A., & Urbina, S. (1997). *Psychological testing.* Upper Saddle River, N.J. USA: Prentice Hall, 7th international edition.

Australian Bureau of Statistics. (1990). *Socio-economic indices for areas SEIFA*. Canberra, Australian Capital Territory: Australian Government Printing Service.

Baltes, P. B., Reese, W., & Nesselroade, J. R. (1977). *Life-span developmental psychology: Introduction to research methods.* Monterey, CA: Brooks/Cole.

Bjorklund, D. F. (1995). *Children's thinking: Developmental function and individual differences*. Sydney, Australia: Brooks/Cole.

Bornholt, L. J. & Piccolo, A. (in press) Individuality, belonging and children's self concepts: A Learning Spiral Model of self evaluation, performance and participation in physical activities, *Applied Psychology An International Review*.

Bornholt, L. J. (1997). Aspects of Self Knowledge about activities with young children. *Every Child*, *3,* 8-9.

Bornholt, L. J. (2000). Social and personal aspects of self knowledge: A balance of individuality and belonging. *Learning & Instruction*, *10*, 415-429.

Bornholt, L. J. (2005a) *ASK-KIDS Self Concept Inventory. Test and Test Manual*. Melbourne: ACER Press.

Bornholt, L. J. (2005b) *Response of Motivation Spiral Models (MSM) to context: Developing children's motivation in reading, movement and social activities*. Paper presented at the Australasian Human Development Association in Perth in July 2005.

Bornholt, L. J. (2005c) Aspects of self knowledge about activities: An integrated model of self concepts. *European Journal of Psychological Assessment*, *21*, 156-164.

Bornholt, L. J., Ouvrier, R. A., Black, F. H., & Hendy, J. (1999, April). *Social and personal influences on a sense of competence at a cognitive screening test for children*. Presented at the American Education Research Association conference. Montreal.

Bornholt, L. J., Spencer, F. H., Fisher, I. H., & Ouvrier, R. A. (2004). Cognitive screening for young children: Development and diversity in learning contexts. *Journal of Child Neurology*. *19*(5), 313 – 317.

Bornholt, L.J., & Ingram, A. (2001). Personal and social identity in children's self-concepts about drawing. *Educational Psychology, 21* (2), 151-167.

Bowes, J. M., & Hayes, A. (Eds.). (1999). *Children, Families and Communities. Contexts and Consequences.* Melbourne: Oxford University Press.

Brake, N. A., & Bornholt, L. J. (2004). Personal and social bases of children's self-concepts about physical movement. *Perceptual and Motor Skills*, *98*, 711-724.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills, CA USA: Sage.

Coleman, C., & Bornholt, L. J. (2003). Reading self concepts and task choices for children with reading difficulties. *Australian Journal of Learning Disabilities, 8*, 24-31.

Dyche, G. M., & Johnson, D. A. (1991). Development and evaluation on CHIPASAT, an attention test for children: Test-retest reliability and practice effect for a normal sample. *Perceptual and Motor Skills, 72* (2), 563–572.

Eccles, J. S., Wigfield, A., Harold, R. D., & Blumfield, P. (1993). Age and gender differences in children's self and task perceptions during elementary school. *Child Development 64,* 830-846.

Elliot, C. D. (1990). *The Differential Ability Scales*. Marrickville, Australia: Harcourt Assessment Company and the Psychological Corporation.

Ellis, H. C., & Hunt, R. R. (1989). *Fundamentals of human memory and cognition.* Dubuque, Iowa: WC Brown.

Freed, M. N., Ryan, J. M., & Hess, R. K. (1991). *Handbook of statistical procedures and their computer applications to education and the behavioral sciences*. American Council on Education. New York, USA: Macmillan.

Gross, J., & Hayne, H. (1999). Young children's recognition and description of their own and others' drawings. *Developmental Science, 2*, 476-489.

Hannan, T. (2005). Assessing children: Hits and myths. *InPsyc, 27* (3), 14 – 17.

Huck, S.W. (2004). Reading statistics and research, 4th Ed. : New York, USA: Allyn and Bacon.

Kelley-Gomez, D. J. (1999). Test-retest reliability and validity of the computerized version of the category test – young children's version. *Dissertation Abstracts International, 59* (7-9), 2464.

Markham, R., Howie, P., & Hlavacek, S. (1999). Reality monitoring in auditory and visual modalities: Developmental trends and effects of cross-modal imagery. *Journal of Experimental Child Psychology, 72*, 51-70.

Marsh, H. W., Debus, R. & Bornholt, L. J. (2005) Validating Young Children's Self-concept Responses: Methodological Ways and means to understand their responses. (pp. 138-160). In Douglas M. Teti (ed) *Handbook of Research Methods in Developmental Psychology*. Oxford: Blackwell.

Morrison, F. J., Griffith E. M., & Frazier J. (1996). *Schooling and the 5 to 7 shift: A natural experiment. The five to seven year shift: The age of reason and responsibility*. In A. J. Sameroff and M. M. Haith. Chicago, IL, University of Chicago Press: 161-186.

Nelson, K. (1996). *Memory development from 4 to 7 years. The five to seven year shift: The age of reason and responsibility*. A. J. Sameroff and M. M. Haith. Chicago, The University of Chicago Press: 141-160.

Newcombe, N. (1996). *Child development: Change over time*. New York, Harper Collins College.

Ouvrier, R. A., Goldsmith, R. F., Ouvrier, S., & Williams, D. C. (1993). The value of the Mini Mental State Examination in childhood. A preliminary study. *Journal of Child Neurology, 8,* 145-148.

Ouvrier, R., Hendy, J., Bornholt, L. J., & Black, F. H. (1999). The School-Years Screening Test for the Evaluation of Mental Status (SYSTEMS). *Journal of Child Neurology, 14,* 772-780.

Ouvrier, R., Hendy, J., Bornholt, L. J., & Black, F. H. (2000). *The Administration and Scoring Manual for the School-Years Screening Test for the Evaluation of Mental Status (SYSTEMS)*. Westmead, NSW, Children's Hospital at Westmead (available from RobertO@chw.edu.au).

Russell, L., Bornholt, L., & Ouvrier, R. (2002). Brief cognitive screening and self concepts for children with low intellectual functioning. *British Journal of Clinical Psychology, 41*, 93-104.

Sameroff, A. J., & Haith, M .M. (Eds.). (1996). *The five to seven year shift: The age of reason and responsibility.* Chicago, IL: The University of Chicago Press

Sattler, J. M. (2001). *Assessment of Children: Cognitive Applications*. San Diego : J.M. Sattler, 4th edition.

Schuerger, J. M & Witt, A. C. (1989). The temporal stability of individually tested intelligence. *Journal of Clinical Psychology, 45* (2), 294–302.

Spencer, F. H., & Bornholt, L. J. (2003). A model of children's cognitive functioning and cognitive self-concepts. *Australian Journal of Learning Disabilities*, *8*, 4-9.
Spencer, F. H., Bornholt, L. J., & Ouvrier, R. A. (2003). Test reliability and stability of children's cognitive functioning. *Journal of Child Neurology, 18*, 5 – 11.
Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *The Stanford Binet Intelligence Test*. Marrickville, NSW, Harcourt Assessment Company & the Psychological Corporation.
van den Burg, W., & Kingma, A. (1999). Performance of 225 Dutch school children's on Rey's Auditory Verbal Learning Test (AVLT): Parallel test-retest reliabilities with an interval of 3 months and normative data. *Archives of Clinical Neuropsychology, 14*(6), 545–559.
Vassar Statistics Web page http://faculty.vassar.edu/lowry/rdiff.html

Acknowledgements

We are grateful to the insightful children, and to their parents, teachers and principals for their continued co-operation in the project. We would like to thank Clinical Professor Robert Ouvrier at The Children's Hospital at Westmead in Sydney for his support and encouragement. This research was part of a PhD thesis completed by Fiona Spencer (nee: Black). The project was funded by a Public Health Postgraduate Research Scholarship from the National Health and Medical Research Council of Australia.

Notes on Authors

Dr Fiona Spencer (nee Black) (BA *UNE*, GradDipPsych *Canb*, GradDipEd *UNE*, PhD *Sydney*) is a researcher and lecturer in educational psychology and human development at Queensland University of Technology. Her research examines psychological resilience, university student learning and assessment, children's cognitive functioning and self concepts. She supervises research students in working memory and its relationship to cognitive functioning and school achievement, conceptions of learning and values in education, personal responsibility, defining the nature and outcomes of Australian professional supervision, individual learning within the context of working organisations, and ability grouping for mathematically gifted adolescent boys, has published in peer reviewed as well as professional journals, and presented at national and international conferences.

Dr Laurel Bornholt (BAHons *Melb*., PhD *Macq*., *MAPS*) is a researcher and lecturer in social-developmental psychology at the University of Sydney, and holds honorary positions in the School of Psychology and in Psychological Medicine at the Children's Hospital at Westmead. Her research examines aspects of identity (including individuality, belonging, identity of place) in dynamic contexts, and develops innovative assessments of self-concepts, for children, adolescents and adults, in clinical education, community and workplace settings. She has published over 30 research papers in peer reviewed journals, and regularly presents papers at national and international conferences.